Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 485174, 16 pages
doi:10.1155/2012/485174

Review Article

Mixed Modeling with Whole Genome Data

Jing Hua Zhao and Jian'an Luan

MRC Epidemiology Unit & Institute of Metabolic Science, Addrenbrooke's Hospital, Box 285, Hills Road, Cambridge CB2 0QQ, UK

Correspondence should be addressed to Jing Hua Zhao, jinghua.zhao@mrc-epid.cam.ac.uk

Received 2 March 2012; Accepted 20 April 2012

Academic Editor: Yongzhao Shao

Objective. We consider the need for a modeling framework for related individuals and various sources of variations. The relationships could either be among relatives in families or among unrelated individuals in a general population with cryptic relatedness; both could be refined or derived with whole genome data. As with variations they can include oliogogenes, polygenes, single nucleotide polymorphism (SNP), and covariates. *Methods*. We describe mixed models as a coherent theoretical framework to accommodate correlations for various types of outcomes in relation to many sources of variations. The framework also extends to consortium meta-analysis involving both population-based and family-based studies. *Results*. Through examples we show that the framework can be furnished with general statistical packages whose great advantage lies in simplicity and exibility to study both genetic and environmental effects. Areas which require further work are also indicated. *Conclusion*. Mixed models will play an important role in practical analysis of data on both families and unrelated individuals when whole genome information is available.

1. Introduction

Genomewide association studies (GWASs) have successfully identified many genetic variants consistently associated with human diseases or other traits. Both unrelated individuals in a population or related individuals in families have been involved in such studies. There is a variety of issues which merit further consideration.

Our concern here is on correlations among individuals, which are "the central piece of information" [1] in detection and characterization of gene-trait association. Consideration of these correlations has traditionally limited to family data whose critical role in genetic epidemiological study ranges from familial aggregation, segregation, linkage to association [2], and special attention is required in the analysis compared to unrelated individuals from a population. Correlations arise naturally among relatives but can be relevant to

population-based study as well given that relatedness can also be established among unrelated individuals based on whole-genome data in GWASs [3]. The correlations are linked to a long attempt to model influence of multiple genes on a specific phenotype. Specifically, Fisher [4] assumed that a quantitative trait results from many genes with variable small to moderate effects. Concrete evidence of multiple genetic influence has been revealed by recent waves of GWASs on height [5], blood pressure [6], lipids [7], obesity [8], and so forth, leading to the note in [9]. Gene-environment interaction and common environment can be considered similarly.

There is a relativelysmall literature in human genetics to iterate mixed models to account for heterogeneity among groups of individuals compared to the general statistics literature where genetic applications have been acknowledged [10, pages 190–192] and [11, pages 4864–4871]. This is likely due to the complexity with a generic implementation. We therefore conduct a survey of the framework with exploration of general software environments. As will be seen below, it readily applies to human genetics when correlations within these groups are explicitly modeled. The familiar form accommodate effects of major or oligogenes, polygenes, common environment, and unique environment, which collectively contribute to variance of the trait and known as "variance component models" [12, 13]. For instance, individuals' body weight (kg) divided by height^2 (m^2), referred as body mass index (BMI, kg/m^2) and commonly used as surrogate of obesity, varies with the broad heritable background of individuals (polygenes), sex, age, family membership, susceptible genes such as *FTO* [14] (which has a major effect serving an example of oliogene), where sex and age can be considered as fixed effects while variability attributable to (expected) correlation between members of family as with *FTO* are random effects [15–18]. The flexibility of such a framework may be missing in various computer programs (see http://linkage.rockefeller.edu). As for outcome of interest, it is usually quantitative or binary traits, with [19] as an exception. The implementation we consider will be *SAS* (see http://www.sas.com) [20] and *R* (see http://www.r-project.org) [21] with a Cox model counterpart [22]. A note on Bayesian counterpart is also ready [23–25], especially for linkage [1], association [26] and implementation in *Morgan*. To save space, we consequently omit reference to programs when they are available from the lists given here.

We attempt to connect various models in our survey paying special attention to their use in data analysis. We show that with generic facilities as available from *R*, we can accommodate additional outcomes such as count, survival, as well as account for information such as identity-by-descent (IBD) or common environment. We will illustrate with the family data available to genetic analysis workshops (GAWs) (see http://www.gaworkshop.org) 16 and 17. We will also discuss the implications of whole genome data availability via connection to earlier literature.

2. Models

As will soon become clear, the framework is essentially motivated from the usual general linear model (GLM) or generalized linear mixed model (GLMM) allowing for correlated random effects, including the Cox regression model. We will briefly describe the models as an analogy between GLM and GLMM but will not go into details of their estimation procedures, as both are widely available.

2.1. GLM

We start from the usual GLM disregarding familial correlations. Let the phenotypes of n individuals in a family be $(y_1, \ldots, y_n)$, its distribution is exponential

$$f(y_i, \theta_i, \varphi) = \exp\left[\frac{y_i\theta_i - b_i(\theta_i)}{\varphi} + c(y_i, \varphi)\right], \tag{2.1}$$

where $b(\cdot)$ and $c(\cdot)$ are known functions, φ a scale or dispersion parameter. Furthermore, let $E[y_i] = \mu_i$ and let this be connected to a linear predictor using link function $g(\cdot)$ by $\eta_i = g(\mu_i) = X_i\beta$, where X_i is a vector of covariates and β the regression coefficient(s). For simplicity, only canonical link is used so that $\theta_i = \mu_i$. It can be shown [27] that the expectation $E(y_i) = \mu_i = b'(\theta_i)$ and variance $V(y_i) = \varphi b''(\theta_i)$. Some special cases as with their properties are well-recognized [28], for which models involving continuous and binary outcomes are most common.

Normal: $y_i \sim N(\mu_i, \sigma_i^2)$, we have $\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2$, $\varphi = \sigma_i^2$, $b'(\theta_i) = \theta_i$, $\varphi b''(\theta_i) = \sigma_i^2$ and an identity link.

Binomial: $y_i \sim \text{Binom}(n, \mu_i)$, $\theta(\mu_i) = \ln(\mu_i/(1-\mu_i))$, $b(\theta_i) = \ln(1+\exp(\theta_i))$, $\varphi = 1/n$, $b'(\theta_i) = \exp(\theta_i)/(1+\exp(\theta_i))$, $\varphi b''(\theta_i) = \mu_i(1-\mu_i)/n$, and a logit link $g(\mu_i) = \ln(\mu_i/(1-\mu_i))$.

Analysis of censored survival data can be molded into the framework [29]. Let t_i denote the event time, c_i the censoring time and $\delta_i = I(t_i \le c_i)$ the event indicator for unit $i, i = 1, \ldots, n$; the basic Cox model with vector of explanatory variables X_i is specified via a hazard function $\lambda_i(t) = \lambda_0(t)\exp(X_i\beta)$, where $\lambda_0(t)$ is the baseline hazard function. The partial likelihood (PL) for the standard Cox model can be expressed as follows:

$$\text{PL}(\beta) = \prod_{i=1}^{n}\left[\frac{\exp(X_i\beta)}{\sum_{j\in R(t_i)}\exp(X_j\beta)}\right]^{\delta_i}, \tag{2.2}$$

where n failure times have been ordered such that $t_1 < \cdots < t_n$ and $R(t_i)$ is the "risk set," the number of cases that are at risk of experiencing an event at time t_i.

Although GLM lays the foundation in many applications of general statistics, it largely serves a motivating role for models that are capable to account for familial correlations. As shown below, this is achieved with introduction of (correlated) random effects as in GLMM, but it is also linked with other models.

2.2. GLMM

We now consider model involving individual i, $i = 1, \ldots, N$, where N is the total number of individuals in our sample.

Polygene

Let P denote the polygene representing independent genes of small effect, which follows a multivariate normal distribution with covariance matrix

$$g(\mu_i) = X_i\beta + P_i. \tag{2.3}$$

The likelihood for all relatives is furnished with specification of the distribution of $P = (P_1, \ldots, P_N)$ with covariance

$$\Sigma_P = 2\Phi\sigma_P^2, \tag{2.4}$$

where $\Phi \equiv \{\phi_{ij}\}_{n\times n}$ and ϕ_{ij} is the kinship coefficient, defined such that, given two individuals, one with genes (g_i, g_j) and the other with genes (g_k, g_l), the quantity is $(1/4)(P(g_i \equiv g_k) + P(g_i \equiv g_l) + P(g_j \equiv g_k) + P(g_j \equiv g_l))$, where $\equiv$ represents probability that two genes sampled at random from each individual are IBD. The kinship coefficients for MZ twins, DZ twins/full-sibs, parent-offspring, half-sibs, and unrelated individuals are 0.5, 0.25, 0.25, 0.125, and 0, respectively.

The likelihood function for model (2.3) has the following form:

$$L(y_1, \ldots, y_N) = \int L(y \mid P)L(P)dP, \tag{2.5}$$

where $L(y \mid P) = \prod_{i=1}^{N} f(y_i \mid P)$ and $L(P) = \left(\sqrt{2\pi|\Sigma_P|}\right)^{-1} \exp[-P'\Sigma_P^{-1}P/2]$ only involve with random effects, noting that it is assumed that, given random effects in the model, the phenotypic values among n relatives are independent and that the parameters of interest in (2.4) are the variances involving polygene (σ_P^2). Regarding the statistical inference of random effects, since the parameter under the null hypothesis is on the boundary of the parameter space, the test for a specific $\sigma_k^2 = 0$, likelihood ratio statistic testing for the hypothesis that $H_0 : \sigma_P^2 = 0$ versus $H_A : \sigma_P^2 > 0$, is referred to a $0.5\chi_0^2 + 0.5\chi_1^2$ distribution or a score statistic as outlined in [11, 19, page 2961].

Oligogene

Suppose that a major gene M is also involved, independently and normally distributed with mean 0 and variances σ_M^2, then the covariance matrix has the form

$$\Sigma_M = \sigma_M^2\Pi, \tag{2.6}$$

where $\Pi \equiv \{\pi_{ij}\}_{N\times N}$ in which π_{ij} is the proportion of alleles shared (IBD) at the major gene between relatives i and j which can be estimated from a multipoint data, so that when it acts additively with polygene P, the likelihood is furnished with an extended covariance

$$\Sigma_{M,P} = \Sigma_M + \Sigma_P. \tag{2.7}$$

For a test of a strictly positive variance associated with a polygene versus polygene and an oligogene, the log likelihood ratio test statistic is referred to $0.5\chi_1^2 + 0.5\chi_2^2$ [30].

Multiple Random Effects

The framework in (2.3) includes the common distributions such as normal, gamma, binomial and Poisson as special cases. For simplicity, we consider a quantitative trait, whose probability density function is normal and a statistical model is as follows:

$$y = X\beta + U + \epsilon, \tag{2.8}$$

and $U \sim N(0, \Sigma)$, $\epsilon \sim N(0, \sigma^2)$, $\mathrm{Cov}(U, \epsilon) = 0$. The expression of Σ^{-1} relative to the precision $1/\sigma^2$ of ϵ as a Cholesky factorization $\Delta'\Delta$, that is, $\Sigma^{-1}/(1/\sigma^2) = \Delta'\Delta$ led to the term *relative precision factor* for Δ [31]. Note that the partition of effects as being fixed and random (H_A: genetic effect) can be compared to a sporadic model (H_0: no genetic effect) $y = X_1\beta_1 + X_2\beta_2 + e$, where both β_1 and β_2 are fixed effects, the involvement of Σ or more specifically Σ^{-1} as a "ridge factor" creates shrinkage in the random effects solutions to the normal equations, that is, "regression towards the mean."

We will see an example from the GAW17 data below that a quantitative trait Q1 is influenced by polygenic background and specific gene *VEGFC* as captured by kinship or relationship matrix and IBD matrix, respectively. This prompts the need to consider multiple random effects. We therefore pursue (2.8) further. As in [32], write $y = X\beta + Z_1 a_1 + \cdots + Z_k a_k + \epsilon$ with the usual assumption that y is $N \times 1$ vector of observations, X an $N \times p$ known matrix, not necessarily of full column rank, β a vector of fixed effects, Z_i a known $N \times r_i$ matrix of rank r_i, a_i random effects with $E(a_i) = 0$, $\mathrm{cov}(a_i) = \sigma_i^2 I_{r_i}, \mathrm{cov}(a_i, a_j) = 0, i \neq j, \mathrm{cov}(a_i, \epsilon) = 0, i, j = 1, \ldots, k$, ϵ an $N \times 1$ vector of errors with $E(\epsilon) = 0, \mathrm{cov}(\epsilon) = \sigma^2 I_N$. Then $E(y) = X\beta$ and $\mathrm{cov}(y) = \Sigma = \sigma^2 I_N + \sum_{j=1}^{k} \sigma_j^2 Z_j Z_j'$. This turns out to be critical to explore the covariance structure involving more (k) parameters $(\sigma_1^2, \ldots, \sigma_k^2)$ in the form

$$\sum\left(\sigma_1^2, \ldots, \sigma_k^2\right) = \sum_1\left(\sigma_1^2\right) + \cdots + \sum_k\left(\sigma_k^2\right), \tag{2.9}$$

where $\Sigma_i(\sigma_i^2)$ has the form of $\sigma_i^2 H_i$, $i = 1, \ldots, k$ with σ_i^2 being the unknown parameter and H_i a (known) coefficient matrix. It will also hold when different variance components such as multiple major genes of interest, gene-gene, gene-environment interactions, common shared environment are to be modeled. For significance test, Case 4 in [30] serves as a general guideline.

A closely related model is the so-called *marginal or population-average model* whereby familial relationship can be specified for e, namely, generalized estimating equations (GEEs) [12, 33]. Given $\mu_i = E(y)$, $V_i = \mathrm{Var}(y)$, it has the form

$$\sum_i \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1}(y_i - \mu_i) = 0, \tag{2.10}$$

for which only link function and variance need to be specified. Parameter estimates are consistent even when variance structure is misspecified, but the ability to use (2.9) is an apparent advantage.

We now turn to the Cox model. First, the consideration of an unobserved family specific random effect is often termed as frailty model, such that families with a larger value

of the frailty will experience the event at earlier times and most "frail" individuals will fail early [34]. Now we allow for correlated frailty and, in analogy to model (2.3) and [22], the appropriate model with random effect U_i becomes $\lambda_i(t) = \lambda_0(t)\exp(X_i\beta + U_i)$. Assuming the parameters of interest are β and σ^2 we have

$$\mathrm{PL}(\beta, U) = \prod_{i=1}^{N}\left[\frac{\exp(X_i\beta + U_i)}{\sum_{j\in R(t_i)}\exp(X_j\beta + U_i)}\right]^{\delta_i}. \tag{2.11}$$

The so-called integrated log likelihood is derived as

$$L = \int \mathrm{PL}(\beta, U)L(U)dU. \tag{2.12}$$

A more tractable solution is via a Laplace approximation for an approximate marginal log likelihood that can be maximized by a penalized partial likelihood (PPL) for parameters (β, σ^2), $\mathrm{PPL}(\beta, U) = \log(\mathrm{PL}(\beta, U)) - U^T\Sigma^{-1}U/2$, followed by a profile likelihood function involving only σ^2.

Furthermore, we can take advantage of the generic form of covariance in other types of models as well. A straightforward yet remarkably useful extension is the multivariate model. For instance, consider (2.8) with m phenotypes. Let $y = (y_{11}, \ldots, y_{1N}, \ldots, y_{mN})^T$ be a vector of m multivariate phenotypes for N individuals. Let β be a vector of dimension mp of the regression coefficients for the p covariates including a vector of 1's corresponding to the overall mean, $X = I_m \otimes X_{N,p}$, an $mN \times mp$ known matrix of covariate values. An analogy to (2.7) and (2.8) lead to the variance-covariance matrix of the m phenotypes with dimension $mN \times mN$ is

$$\Sigma = A \bigotimes \Pi + B \bigotimes R + C \bigotimes I, \tag{2.13}$$

where R is the $N \times N$ matrix of the coefficients of relationship, Π is an $N \times N$ matrix of estimated proportion of alleles IBD, and A, B, C are oligogenic, polygenic, and residual variance-covariance matrices each with dimension $m \times m$.

2.3. Meta-Analysis

One indispensable element in current GWASs is meta-analysis, typically involving findings from both unrelated individuals in a population and those from family data. While we have seen that mixed models are appropriate for a variety of traits in family-based association studies, broadly models for meta-analysis also fall into the same framework as described above. One can imagine a meta-analysis involving individual participant data (IPD). A good summary of approaches for IPD meta-analysis is available [35].

In the two-step approach, the individual participant data are first analysed in each separate study independently by using a statistical method appropriate for the type of data being analysed; for example, a linear regression model might be fitted for continuous responses such as blood pressure, or Cox regression might be applied for time to event data. (This step produces aggregate data for each study including effect estimate and its standard error). These data are then synthesised in the second step using a suitable model

for meta-analysis of aggregate data, such as one that weights studies by the inverse of the variance while assuming fixed or random effects across studies. In the one-step approach, the individual participant data from all studies are modelled simultaneously while accounting for the clustering of participants within studies. This approach again requires a model specific to the type of data being synthesised, alongside appropriate specification of the assumptions of the meta-analysis (e.g., of fixed or random effects across studies).

The two-step approach is the usual one used in various GWAS consortia while a one-step approach for all studies in our context could involve unrelated population-based samples and family data in the meta-model as long as the correlation structure is appropriately specified. The practicality of both approaches has been illustrated in the literature [36, 37] but, in view of the complexity involving in such a framework, and the practical difficulty that a researcher may not have access to individual data from all studies, we refrain ourselves from such a consideration for now but remain focusing on family data as illustrated with both simulated and real data.

2.4. Related Results and Implementations

There have been concerns in the literature regarding large number of units each with bounded size [38] and a large number of random effects [39]. In our context large number of families, each with bounded members, consistent estimate of the random effect is difficult to obtain though fixed effects and variance components will be consistent. However, Type I error rate and power have been explored before [19, 22, 26, 40], so there will be more on specific examples.

Instead of using purposely written programs, we chose to use *R*, for its wide availability and many other features [41], and in particular procedures to fit models described earlier are to a great extent available, including generic procedures from *nlme*, *lme4,* and *gee*, among others, but package designed for family data is *pedigreemm* with *lmekin* for linear mixed models available from *coxme*. We will also compare them to *SAS*, due to its ability to deal with large data, and great flexibility in model specification.

3. Examples

We consider two examples from GAWs 17 and 16, which involve simulated and real data widely available and allow for a lot of experiments to be done.

3.1. GAW17 Data

Data distributed by GAW17 were based on a collection of unrelated individuals and their genotypes were generated from the 1000 Genomes Project (see http://www.1000genomes.org/), from which a sample of 697 individuals in 8 extended families and their genotypes and phenotypes was available. A total of 202 founders in the family data set were chosen at random from the set of unrelated individuals. Replicates of the trait were generated 200 times, but the simulated genotypes remain constant over replicates. The traits made available were Q1, Q2, Q4, and AFFECTED (coded 0 = no 1 = yes) with covariates AGE and SMOKE. The variables describing family structures were ID, FA, MO, SEX (1 = men, 2 = women). Fully informative IBD information was available for 3205 genes.

We chose to examine traits Q1, Q2, and AFFECTED as representatives of quantitative and qualitative traits. According to [42], vascular endothelial growth factor (*VEGF*) pathway was enriched and here vascular endothelial growth factor C (*VEGFC* (see http://en.wikipedia.org/wiki/Vascular_endothelial_growth_factor_C)) was chosen as a causal variant associated with Q1 but not Q2. Q1 also increased with age, and the fact that AFFECTED is a function of Q1 offers the possibility to furnish a logistic regression model and explore age at onset via a Cox model. For illustration, we used age as surrogate for age onset. Being aware of the fact that this was only an approximation, whenever multipe affected individuals within a sibship are available, their average age was used. Causal variants and associate genes provide information on power of association testing statistics while the noncausal counterparts provide analogous results on Type I error rate.

The statistical significance was assessed according to log likelihood ratio tests between models using relationship only versus using both relationship and IBD information. The computation for this is relatively fast; results for all 200 replicates took 1 hour and 48 minutes on our 20-node Linux clusters each with 16 GB RAM and 4 CPUs using Sun grid engines. The nominal significance levels are shown in Table 1, which reveal that the tests are both close to the expected levels under H_0 and H_A.

Gene-based analysis was also conducted for Q1 involving all 3205 genes and the results are shown with selected candidates highlighted in Figure 1, which agree with the simulated model in which the significant regions were in VEGFC/VEGFA.

As one would be keen to see various parameter estimates in a real analysis, we also provide results associated with replicate one. Q1 as based on restricted maximum likelihood (REML) is shown in Table 2. The models with relationship only and with both relationship and IBD information have −2 Res(tricted) log likelihood being 1789.5 and 1775.2, respectively while Akaike Information Criteria (AIC) being 1793.5 and 1781.2, respectively so that using IBD information improved fit for Q1 (smaller AIC). For AFFECTED the results based on maximum pseudolikelihood are shown in Table 3 and those from Cox model in Table 4. Note that the improvement in terms of −2 log pseudolikelihood from 3434.4 to 3445.7 was also substantial. To explore the multivariate model (2.13) involving the polygenic effects for Q1, Q2, and Q4, the six parameters (σ_{11}, σ_{21}, σ_{22}, σ_{31}, σ_{32}, σ_{33}) in the variance-covariance matrix have been expressed according to (2.9). The appropriate matrices associated with all parameters are constructed a priori. These are then subject to procedures such as PROC MIXED and *lmekin*. The joint model of Q1, Q2, Q4 is shown in Table 5.

The implementations are provided in Supplementary information available online at doi: 10.1155/2012/485174. While code blocks shown there are appropriate for one instance, it is preferable to use *SAS*'s output delivery system (ODS) to save various results into databases.

3.2. The Framingham Heart Study

The Framingham Heart Study is under the direction of National Heart, Lung, and Blood Institute (NHLBI) which began in 1948 with the recruitment of adults from the town of Framingham, Massachusetts. Data available for GAW16 were 7130 individuals from the original cohort (373), the first generation cohort (2760), and the third generation cohort (3997) with sex, age, height, weight, blood pressure, lipids, smoking, and drinking. Data as outlined in [43] was used here, where 6848 had genotype data for at least one of the four specified SNPs (rs1121980, rs9939609, rs17782313, and rs17700633). Data for 96 individuals without

any phenotype data but with genotype data and an additional 227 individuals without being assigned a family ID were excluded from analyses. Additionally, four individuals had no data on weight; 86 observations were measured at <18 years of age, and therefore were excluded. The 6,520 remaining individuals were part of 962 families, among which 2073 individuals had completed four visits. Meanwhile, there were also 365 cases of diabetes with their ages of onset.

Kinship information was obtained from family structure and used for genotype-trait association. Computer program *PLINK* [44] with the *−genome* option was also used to infer correlations ($\widehat{\pi}$) using whole genome data. A total of 8485 SNPs on Affymetrix 500 K chips were derived from a panel of 45620 informative autosomal SNPs used in our consortium analysis. This led to estimates for $6520(6520-1)/2 = 21251940$ pairs of relationship. The genetic distance according to $|\pi - \widehat{\pi}|$ [45], that is, sum(abs*(EZ-PI_HAT), na.rm=TRUE)*, is 3421.724. Approximately half (10478474) had $\widehat{\pi}$ of 0.01 or more. Although there was a good agreement between kinship according to the specified family structures and $\widehat{\pi}$, 11207 pairs of individuals deemed to be unrelated had $\widehat{\pi}$ between 0.1–0.3 and 12 of which were greater than 0.3.

Both types of relationship matrices were used for the Cox model via *kinship* and *bdsmatrix.ibd* functions in *R*. The frailty and polygenic models had log likelihoods of −1788.53, −1791.93 with variance estimates 0.10^2 and 0.02^2, respectively. However, with inferred relationship the log likelihood turned out to be −1762.69 and variance estimate 0.24^2. Similar model for BMI at wave 1 was also fitted; a family specific random intercept model yielded log likelihood of −19273.26 and variance 3.44, while a correlated random intercept model gave log likelihood −19379.3 and variance 0.01^2 with comparable results from inferred relationship though with a smaller residual error. The results on diabetes might have suggested a substantial genetic effect while for BMI the use of inferred relationship performed equally well with a model using explicit family structures.

4. Discussion

The models we have considered extend counterparts for unrelated sample by taking into account correlation within and heterogeneity between families. To a large extent, we have presented an appreciation of models and implementations for related individuals using mixed models. At the meantime, we have envisaged a whole range of analyses that can be put in the framework. However, compared to [13] and especially [19], our development is more incremental and helps to gain insight into more complicated models. As a key feature of the model specification, oligogenes, polygenes, common environment, gene-environment interaction, and multivariate data are accommodated in a coherent framework via appropriate covariance structure. The generic nature has enabled a range of genetic association studies. Our interpretation of the model also naturally extends the model for quantitative traits outlined by [19, 46]. It has been recognized that for longitudinal data some commonly used covariance structures, such as compound symmetry, can be expressed as "linear covariance of dimension k" [47, page 258]. Although it could be more involved, it may be possible in our context. Data as in consortium meta-analysis analysis is also perceived in broader framework consisting of both unrelated and related individuals.

We should be aware that mixed models are quite general and may well be linked to other models. For instance, we noticed that model (2.10) is reminiscent of an approach proposed for generalized method of moments [48]. An example as with its link with

individual empirical Bayes estimates has been provided by [49, 50]. A reviewer has brought to our attention recent work on nonparametric methods for longitudinal data [51] and the utility of mixed models in controlling for bias of population stratification (e.g., [52]). This paper has limited coverage of literature on longitudinal analysis of family data, mainly owing to the fact that there is greater difficulty in implementation via general software package. However, this is expected to change. To our knowledge, little work has been done on joint analysis of individual data in the GWAS meta-analysis context. In view of the popularity of consortium data analysis, it will be appealing to have the appropriate mechanism to make it possible.

The models and their implementations are connected with whole genome data in several ways. First, the transition from the variance components models in earlier literature becomes more explicit. More specifically, the models described here are appropriate for GWAS where genetic variants coupled with a high resolution map are available. In general, the variance component associated with a major gene as in (2.7) is a function of the recombination rate (r) [12], that is, $\sigma_M^2 f(r, \pi_{ij})$, where π_{ij} represents identity-by-descent sharing between a pair of individuals i, j for the marker locus; with dense marker, we can assume that $r = 0$ which is also true with (2.9). Second, as in the Framingham data there is a further benefit with dense genetic markers such that they can be used to infer family structure [53] or (global) IBD information [54]. The availability of the deep sequencing data and a long list of established genes are likely to give greater weight on use of family data [55]. It is also desirable that cryptic relatedness in population-based sample can be appropriately taken into account in association analysis. In our own EPIC-Norolk GWAS, samples with cryptic relatedness have been excluded at the quality control stage [56]. It is interesting to note that *coxme* was developed for handling large pedigrees involving sparse matrices; the availability of whole genome data will alter the scenario slightly but nevertheless remain in the same framework. Third, more work is required to shorten computing time. In the literature, it has been proposed to absorb the relationship in the model for quantitative trait by multiplying inverse of the kinship matrix followed by a linear regression, or using residuals from a phenotype-covariate only regression as outcome in a model including SNPs as in *GenABEL*. In principle one can extend the idea to multivariate or longitudinal models where the residuals are obtained only once for GWAS or incorporating regional information before turning to SNP-specific analysis. There are also alternative approaches such as retrospective methods found in *Merlin*. With its greater requirement in computation the "measured genotype" approach here remains intuitive especially for gene-environment characterization. To this point, associate projects such as *BORDICEA* (see http://www.srl.cam.ac.uk/genepi/boadicea/boadicea_home.html) and *BayesMendel* (see http://bcb.dfci.harvard.edu/BayesMendel/) have contributed to the success of work on *R* described here.

A reviewer has expressed interest regarding the Type I error linking to results shown in Table 1. We believe that data as distributed by GAW17 as they were (200 replicates) are not ideal for assessing Type I error and possibly require a bootstrap procedure. In general, from our experience (and personal communications with Profs. Douglas Bates and Terry Therneau), this is a difficult issue and possibly problem specific. In fact, in the recent implementation of GLMM in *lme4*, the associate p values for fixed effects are not shown which nevertheless may leave users with temptation to employ normal approximation. Although we have not conducted extensive numerical experiments, results from GAW17 and the Framingham Study have indicated good performance of these models, and that of the inferred relationship based on whole genome data is impressive. Since only directly

Table 1: Nominal significance according to *VEGFC*.

	Q1	Q2	AFFECTED
Significance level	Power	Type I error	Power
.05	.989	.060	.880
.01	.907	.016	.730
.001	.665	0	.555
.0001	.412	0	.420
.00001	.225	0	.305
.000001	.104	0	.200

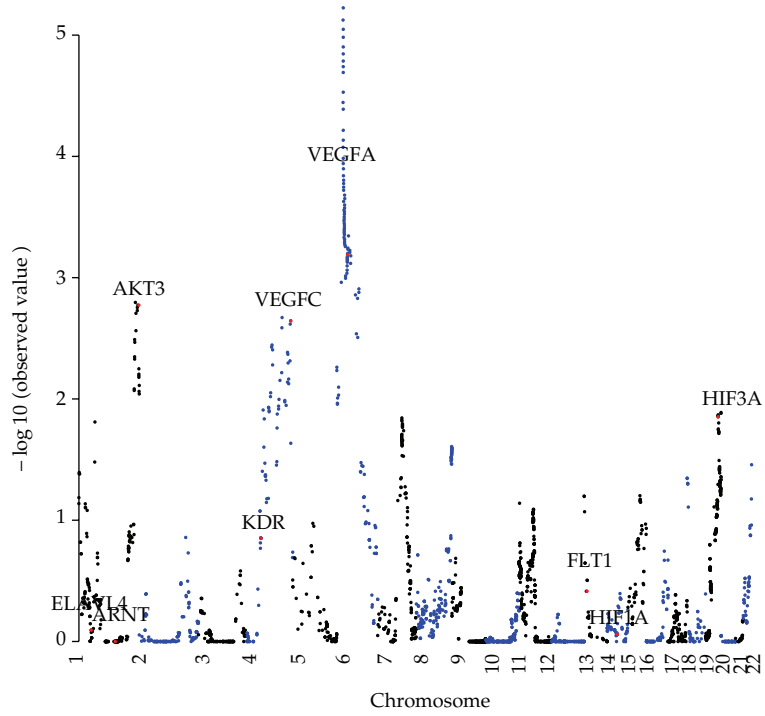

Figure 1: Manhattan plot of Q1 and IBD information where the true loci are highlighted.

genotyped Affy500K SNPs were used, the addition of imputed genotypes, say based on the HapMap, should help to improve the inference. Its use in the usual genomewide association analysis should be considered.

Our attention lies on the implementation by taking advantages of the available implementation in general statistical computing environment. The clarification of the implementation in these should facilitate practical analysis of family data. Although these models are conceptually simple, availability of their implementation vary, notably the ability to allow for both oligogenes and polygenes in a GLMM framework. For *R*, these are at least possible with *nlme*, *lme4*, and additionally *coxme*. At the moment, applications of packages in *R* are often restricted with *lmekin* in *coxme* offering outcomes only on continuous outcome but for *pedigreemm* it is unable to handle complex covariance structure. It is desirable that a function called *nlmekin* can be developed as with *pedigreemm* expanded to incorporate

Table 2: Q1 and *VEGFC* under a linear model.

Model/parameter	Estimate	SE	$z/t^{\dagger}$	−2 Res log likelihood	AIC
Kinship				1789.5	1793.5
σ_P^2	0.5488	0.08262	6.64		
SEX	−0.2379	0.04614	−5.16		
AGE	0.01014	0.001345	7.54		
SMOKE	0.36894	0.07280	5.07		
Kinship + IBD				1775.2	1781.2
σ_P^2	0.4157	0.08713	4.77		
σ_M^2	0.1076	0.03846	2.80		
SEX	−0.2488	0.04542	−5.48		
AGE	0.01044	0.001334	7.82		
SMOKE	0.3821	0.07181	5.32		

$^{\dagger}z$ is for variance components while t for fixed effects.

Table 3: AFFECTED and *VEGFC* under a logistic model.

Model/parameter	Estimate	SE	t	−2 log pseudolikelihood
Kinship				3434.4
σ_P^2	1.3170	0.4376		
SEX	−0.00822	0.2042	−0.04	
AGE	0.07181	0.006047	11.87	
SMOKE	0.9098	0.2285	3.98	
Kinship + IBD				3445.7
σ_P^2	0.6918	0.5989		
σ_M^2	0.4868	0.3698		
SEX	0.006923	0.2048	0.03	
AGE	0.07211	0.006114	11.79	
SMOKE	0.9429	0.2290	4.12	

Table 4: AFFECTED and *VEGFC* under a Cox model.

Model/parameter	Estimate	SE	z	Integrated/penalized likelihoods†
Kinship				−998.8/−980.6
σ_P^2	0.2073			
SEX	0.05267	0.1541	0.34	
SMOKE	0.5000	0.1622	3.08	
Kinship + IBD				−996.1/−967.3
σ_P^2	0.002690			
σ_M^2	0.3615			
SEX	0.07146	0.1603	0.43	
SMOKE	0.5560	0.1696	3.28	

† The log likelihood under the null is −1003.9.

Table 5: Q1, Q2, and Q4 under a multivariate polygenic model.

	Estimate	SE	Log likelihood
	Linear coefficients		−1393.867
c1	0.565	0.108	
c2	0.531	0.109	
c3	0.526	0.109	
Sex	−0.005	0.043	
Age	−0.013	0.001	
Smoke	−0.019	0.051	
	Variance coefficients		
σ_{11}	4.219	0.227	
σ_{12}	−0.103	0.166	
σ_{22}	4.542	0.244	
σ_{31}	0.601	0.178	
σ_{32}	−0.108	0.183	
σ_{33}	5.115	0.275	

additive covariance structures. *SAS*, MIXED, GLIMMIX, and NLMIXED together provide a rich source of practical modeling functionality though the Cox model counterpart is not available. The tackling of various issues has led to efficient algorithm [25]. When the interest is on correlation between multiple traits, the use of *nlme* for multivariate longitudinal data in unrelated individuals has been described [57]. In general, this could be complicated with longitudinal familial data without [58] or with [59] consideration of relationship. In study of obesity-related traits, *FTO* has been shown to be strongly associated with BMI and supported by cross-sectional data as in [14], longitudinal data as in [43] and data across life span as in [60]. Our previous attempt [43], was based on a three-level model and it would be of interest to use kinship information as well.

While the framework we have outlined is comprehensive, we feel that our "proof of concepts" here awaits for extensive testing. It is also desirable that the current implementation can be optimized in computing time. A lot of work has been done for quantitative genetics in plants and animals. Our experience indicated that the running time with *SAS* was longer time than *R*. However, in an analysis of longitudinal lung function data in the EPIC-Norfolk study, we have shown that although an individual analysis could be slow, it is possible to perform an analysis for GWAS using *SAS* and Linux clusters so that ~2.5 M SNPs would finish within 14 hours when running each chromosome on a separate node. It is likely that it benefited from *SAS* caching frequently-used instructions. Greater proportion of coding in *C/C++* should also be helpful. Given the utility of the popular environments can be shown, their take-up in genomewide association studies will be quick and it is very much in line with efforts in other disciplines where large volume of data is involved.

Acknowledgments

A lot of the insights were gained during analysis of GAWs 14, 16, 17 and in particular maintenance of the *R* counterpart of the *S-PLUS* package *kinship* (http://mayoresearch.mayo.edu/mayo/research/biostat/upload/kinship.pdf) by the first author. The authors are therefore very grateful of the pioneering work and advices given by Profs Terry Therneau,

Beth Atkinson, and Mariza de Andrade all at the Mayo Clinic and interactions with many other colleagues elsewhere. The comprehensive *R* archive network (CRAN (http://cran.r-project.org)) as with Professors Kurt Hornik and Brian Ripley has been a constant source of support. The work presented here was partly done for CompBio2011 and useR!2011. They wish to thank Drs. Qihua Tan, Fuzhong Xue, Wendi Qian, and Luigi Palla for their participation and comments during the GAWs 16 & 17 analysis which led to this work, Dr Wendi Qian's comments on *SAS* PROC GLIMMIX, and Dr Antonis Antoniou's suggestion of using average age within a sibship to approximate age at onset. The example regarding twins was due to a query from Dr. Marcel de van Hoed. They are also grateful of the Editor for communications which led to the work on the paper and three anonymous reviewers for their insightful comments which led to its improvement. The work reported here also allows us for making minor changes to the syntax shown in [36]. Professor Peter McCullagh from University of Chicago and Dr. David Clifford from CSIRO have kindly provided advices regarding the use of *regress*.

References

[1] D. C. Thomas and W. J. Gauderman, "Gibbs sampling methods in genetics," in *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richard, and D. J. Spiegelhalter, Eds., pp. 419–440, Chapman & Hall/CRC, London, UK, 1996.

[2] D. C. Thomas, *Statistical Methods in Genetic Epidemiology*, University Press, Oxford, UK, 2004.

[3] J. Yang, B. Benyamin, B. P. McEvoy et al., "Common SNPs explain a large proportion of the heritability for human height," *Nature Genetics*, vol. 42, no. 7, pp. 565–569, 2010.

[4] R. A. Fisher, "The correlation between relatives on the supposition of mendelian inheritance," *Transactions of the Royal Society of Edinburgh*, vol. 52, pp. 399–433, 1918.

[5] H. L. Allen, K. Estrada, G. Lettre et al., "Hundreds of variants clustered in genomic loci and biological pathways affect human height," *Nature*, vol. 467, no. 7317, pp. 832–838, 2010.

[6] G. B. Ehret, P. B. Munroe, K. M. Rice et al., "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk," *Nature*, vol. 478, no. 7367, pp. 103–109, 2011.

[7] T. M. Teslovich, K. Musunuru, A. V. Smith et al., "Biological, clinical and population relevance of 95 loci for blood lipids," *Nature*, vol. 466, no. 7307, pp. 707–713, 2010.

[8] E. K. Speliotes, C. J. Willer, S. I. Berndt et al., "Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index," *Nature Genetics*, vol. 42, no. 11, pp. 937–948, 2010.

[9] R. Plomin, C. M. A. Haworth, and O. S. P. Davis, "Common disorders are quantitative traits," *Nature Reviews Genetics*, vol. 10, no. 12, pp. 872–878, 2009.

[10] C. E. McCulloch and S. R. Searle, *Generalized, Linear, and Mixed Models*, Wiley Series in Probability and Statistics, Wiley-Interscience, New York, NY, USA, 2001.

[11] SAS Institute, *SAS/STAT 9.3 User's Guide*, SAS Publishing, Cary, NC, USA, 2011.

[12] C. I. Amos, "Robust variance-components approach for assessing genetic linkage in pedigrees," *American Journal of Human Genetics*, vol. 54, no. 3, pp. 535–543, 1994.

[13] J. Blangero, J. T. Williams, and L. Almasy, "Variance component methods for detecting complex trait loci," *Advances in Genetics*, vol. 42, pp. 151–181, 2001.

[14] T. M. Frayling, N. J. Timpson, M. N. Weedon et al., "A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity," *Science*, vol. 316, no. 5826, pp. 889–894, 2007.

[15] N. E. Morton and C. J. MacLean, "Analysis of family resem blance. III. complex segregation of quantitative traits," *American Journal of Human Genetics*, vol. 26, pp. 489–503, 1974.

[16] J. L. Hopper and J. D. Mathews, "Extensions to multivariate normal models for pedigree analysis," *Annals of Human Genetics*, vol. 46, no. 4, pp. 373–383, 1982.

[17] K. Lange and M. Boehnke, "Extensions to pedigree analysis. IV. Covariance components models for multivariate traits," *American Journal of Medical Genetics*, vol. 14, no. 3, pp. 513–524, 1983.

[18] S. J. Hasstedt, "A mixed-model likelihood approximation on large pedigrees," *Computers and Biomedical Research*, vol. 15, no. 3, pp. 295–307, 1982.

[19] M. P. Epstein, J. E. Hunter, E. G. Allen, S. L. Sherman, X. Lin, and M. Boehnke, "A variance-component framework for pedigree analysis of continuous and categorical outcomes," *Statistics in BioSciences*, vol. 1, no. 2, pp. 181–198, 2009.
[20] A. M. Saxton, Ed., *Genetic Analysis of Complex Traits Using SAS*, SAS Publishing, 2004.
[21] A. I. Vazquez, D. M. Bates, G. J. M. Rosa, D. Gianola, and K. A. Weigel, "Technical note: an R package for fitting generalized linear mixed models in animal breeding," *Journal of Animal Science*, vol. 88, no. 2, pp. 497–504, 2010.
[22] V. S. Pankratz, M. de Andrade, and T. M. Therneau, "Random-effects cox proportional hazards model: general variance components methods for time-to-event data," *Genetic Epidemiology*, vol. 28, no. 2, pp. 97–109, 2005.
[23] V. Ducrocq and G. Casella, "A bayesian analysis of mixed survival models," *Genetics Selection Evolution*, vol. 28, no. 6, pp. 505–529, 1996.
[24] D. Sorensen and D. Gianola, *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*, Springer, New York, NY, USA, 2002.
[25] P. Waldmann, "Easy and flexible Bayesian inference of quantitative genetic parameters," *Evolution*, vol. 63, no. 6, pp. 1640–1643, 2009.
[26] P. R. Burton, K. J. Scurrah, M. D. Tobin, and L. J. Palmer, "Covariance components models for longitudinal family data," *International Journal of Epidemiology*, vol. 34, no. 5, pp. 1063–1079, 2005.
[27] J. M. Lachin, *Biostatistical Methods: The Assessment of Relative Risks*, Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ, USA, 2nd edition, 2011.
[28] A. Skrondal and S. Rabe-Hesketh, *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Interdisciplinary Statistics, Chapman & Hall/CRC, Boca Raton, Fla, USA, 2004.
[29] J. Whitehead, "Fitting Cox's regression model to survival data using GLIM," *Journal of the Royal Statistical Society*, vol. 29, no. 3, pp. 268–275, 1980.
[30] G. Verbeke and G. Molenberghs, *Linear Mixed Models for Longitudinal Data*, Springer, New York, NY, USA, 2000.
[31] J. Pinheiro and D. M. Bates, *Mixed Effects Models in S and S-PLUS*, Springer, 2000.
[32] R. B. Bapat, *Linear Algebra and Linear Models*, Universitext, Springer, London, UK, 3rd edition, 2012.
[33] P. J. Diggle, P. J. Heagerty, K.-Y. Liang, and S. L. Zeger, *Analysis of Longitudinal Data*, vol. 25, Oxford University Press, Oxford, UK, 2nd edition, 2002.
[34] J. P. Klein and M. L. Moeschberger, *Survival Analysis-Techniques for Censored and Truncated Data*, Springer, 2nd edition, 2003.
[35] R. D. Riley, P. C. Lambert, and G. Abo-Zaid, "Meta-analysis of individual participant data: rationale, conduct, and reporting," *British Medical Journal*, vol. 340, p. c221, 2010.
[36] J. H. Zhao, J. Luan, R. J. F. Loos, and N. Wareham, "On genotype-phenotype association using SAS," in *Proceedings of the 2nd International Conference on Computational Bioscience*, pp. 428–433, Cambridge, Mass, USA, 2011.
[37] T. D. Pigott, *Advances in Meta-Analysis*, Springer, 2012.
[38] J. Neyman and E. L. Scott, "Consistent estimates based on partially consistent observations," *Econometrica*, vol. 16, pp. 1–32, 1948.
[39] P. Hall, J. S. Marron, and A. Neeman, "Geometric representation of high dimension, low sample size data," *Journal of the Royal Statistical Society*, vol. 67, no. 3, pp. 427–444, 2005.
[40] N. J. Schork, "Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations," *American Journal of Human Genetics*, vol. 53, no. 6, pp. 1306–1319, 1993.
[41] J. H. Zhao and Q. Tan, "Integrated analysis of genetic data with R," *Human Genomics*, vol. 2, no. 4, pp. 258–265, 2006.
[42] L. Almasy, T. D. Dyer, J. M. Peralta et al., "Genetic Analysis Workshop 17 mini-exome simulation," *BMC Proceedings*, vol. 5, article S2, supplement 9, Article ID S2, 2011.
[43] J. Luan, B. Kerner, J. H. Zhao et al., "A multilevel linear mixed model of the association between candidate genes and weight and body mass index using the framingham longitudinal family data," *BMC Proceedings*, vol. 3, article S115, supplement 7, 2009.
[44] S. Purcell, B. Neale, K. Todd-Brown et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.
[45] A. Sanchez, J. Ocaña, and F. Utzet, "Sampling theory, estimation, and significance testing for Prevosti's estimate of genetic distance," *Biometrics*, vol. 51, no. 4, pp. 1216–1235, 1995.

[46] M. de Andrade, E. Atkinson, E. Lunde, C. I. Amos, and J. Chen, "Estimating genetic components of variance for quantitative traits in family studies using the multic," Tech. Rep., Mayo Clinic, 2006.
[47] E. F. Vonesh and V. M. Chinchilli, *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, vol. 154 of *Statistics: Textbooks and Monographs*, Marcel Dekker, NewYork, NY, USA, 1997.
[48] G. Yin, "Bayesian generalized method of moments," *Bayesian Analysis*, vol. 4, no. 2, pp. 191–208, 2009.
[49] T. Moger, O. O. Aalen, K. Heimdal, and H. K. Gjessing, "Analysis of testicular cancer data using a frailty model with familial dependence," *Statistics in Medicine*, vol. 23, no. 4, pp. 617–632, 2004.
[50] O. O. Aalen, O. Borgan, and H. K. Gjessing, *Survival and Event History Analysis: A Process Point of View*, Statistics for Biology and Health, Springer, NewYork, NY, USA, 2008.
[51] Y. Wang, C. Huang, Y. Fang, Q. Yang, and R. Li, "Flexible semiparametric analysis of longitudinal genetic studies by reduced rank smoothing," *Journal of the Royal Statistical Society*, vol. 61, no. 1, pp. 1–24, 2012.
[52] J. Yu, G. Pressoir, W. H. Briggs et al., "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness," *Nature Genetics*, vol. 38, no. 2, pp. 203–208, 2006.
[53] A. G. Day-Williams, J. Blangero, T. D. Dyer, K. Lange, and E. M. Sobel, "Linkage analysis without defined pedigrees," *Genetic Epidemiology*, vol. 35, no. 5, pp. 360–370, 2011.
[54] L. Han and M. Abney, "Identity by descent estimation with dense genome-wide genotype data," *Genetic Epidemiology*, vol. 35, no. 6, pp. 557–567, 2011.
[55] J. R. Lupski, J. G. Reid, C. Gonzaga-Jauregui et al., "Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy," *The New England Journal of Medicine*, vol. 362, no. 13, pp. 1181–1191, 2010.
[56] R. J. F. Loos, C. M. Lindgren, S. Li et al., "Common variants near MC_4R are associated with fat mass, weight and risk of obesity," *Nature Genetics*, vol. 40, no. 6, pp. 768–775, 2008.
[57] S. Bandyopadhyay, B. Ganguli, and A. Chatterjee, "A review of multivariate longitudinal data analysis," *Statistical Methods in Medical Research*, vol. 20, no. 4, pp. 299–330, 2011.
[58] B. C. Sutradhar, *Dynamic Mixed Models for Familial Longitudinal Data*, Springer Series in Statistics, Springer, NewYork, NY, USA, 2011.
[59] J. M. Soler and J. Blangero, "Longitudinal familial analysis of blood pressure involving parametric (co)variance functions," *BMC Genetics*, vol. 4, article S87, supplement 1, 2003.
[60] R. Hardy, A. K. Wills, A. Wong et al., "Life course variations in the associations between *FTO* and MC_4R gene variants and body size," *Human Molecular Genetics*, vol. 19, no. 3, pp. 545–552, 2010.